Big Data Analysis
Key Take Aways

Zertifikatsarbeit im
CAS Big Data Analysis

Zürcher Fachhochschule
HWZ Hochschule für Wirtschaft Zürich

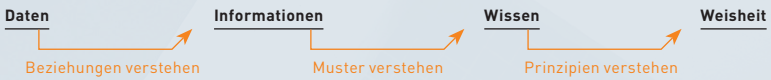
eingereicht bei
Melanie Knobelspies und Umberto Michelucci

vorgelegt von Christian Heimann
Bern, 7. Juni 2019

Editorial

» Big Data

Big Data ist in aller Munde. Noch nie hatte die Menschheit so viele Daten erfasst und zur Verfügung wie heute. Daten, die aus unseren Smartphones stammen, sei es der selbst erzeugte Tweet oder die von Google Maps erzeugten Lokalisierungs-Daten; Daten, die aus Simulationen stammen, welche das Forschen und Experimentieren ein rasantes Tempo annehmen lässt oder auch Daten, die von Systemen erzeugt werden, welche uns tagtäglich den Komfort bieten, den wir als selbstverständlich ansehen – überall sind Daten Teil der Gesellschaft.
Aber aufgepasst, «Daten» sind nicht zu Verwechseln mit «Wissen». Um aus all den Daten «Weisheit» zu schaffen, benötigt es Zwischenschritte, deren Erreichung Aufwände mit sich bringen:



Für die Wirtschaft ist es essenziel, sich mit den vorhandenen Daten zu beschäftigen und sinnvolle Informationen daraus zu generieren. Auf Basis dieser Informationen können Entscheidungen gefällt werden, die nicht nur aus einem Bauchgefühl heraus kommen – wie Sie zum Beispiel bei der Seite zu «Data Driven Organization» sehen werden. Und noch besser, wenn durch die Informationen nicht nur Entscheidungen zu stande kommen, sondern auch gleich noch Aktionen ausgelöst werden können … schnell ist man ganz tief drin, im Thema Big Data Analysis!

» Vorstellungen und Erwartungen

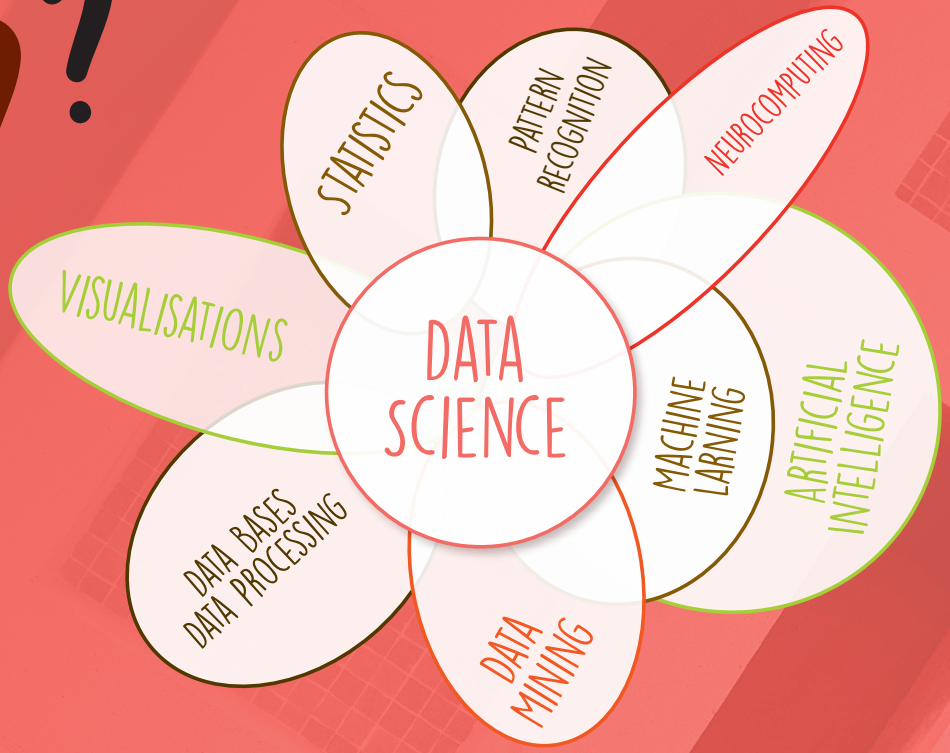
Die Vorstellung, das CAS Big Data Analysis drehe sich stark um den Umgang mit grossen Daten-Mengen und den dazu nötigen Tools und Technologien, wurde nicht erfüllt. Aus berechtigten Gründen. Es wurde klar, dass Schwierigkeiten im Umgang mit grossen Daten-Mengen mit der Technik gelöst werden – und daher sekundär sind. Der Schwerpunkt lag klar auf dem Bereich «Analysis». Analyse kann man auch im kleinen betreiben – man muss dazu aber die Kernkonzepte der statistischen Analyse kennen. Die Vermittlung dieses Wissens war ein wichtiger Teil des CAS, nebst den damit verbundenen, vielseitigen Möglichkeiten deren Nutzung.
Gastreferenten, welche Einblicke in aktuelle Projekte geben konnten, lieferten trotzdem den nötigen Ausblick, um die Brücke zwischen «Analysis» und «Big Data Realität in der Wirtschaft» zu schlagen. Auch diese Inputs wurden wo möglich in die Arbeit aufgenommen.

» Zur Form, zum Zweck

Die Aufgabe dieser Zertifikatsarbeit lautete: Eine spannende Mischung zwischen Präsentation, Lerntagebuch und Nachschlagewerk soll es werden. Für jeden Präsenz-Kurstag mit nachfolgendem Selbststudiums-Tag sollten 2 bis 3 Slides entstehen, welche die wichtigen Erkenntnisse zusammenbringen. Sei es aus der Theorie, schwieriges aus der Programmierung oder grundsätzlich aus dem Ablauf.
Mein Fokus richtete sich weniger auf die Programmierung und mehr auf die konzeptionellen Elemente. Durch den Ansatz, eine Präsentation zu machen – und nicht bloss eine Text-Dokumentation als A4 Hochformat in 10-Punkt-Arial – war klar, dass die visuelle Aufbereitung eine starke Gewichtung bekommen sollte. Das finale Resultat, die «Take Aways» des CAS Big Data Analysis, ist auf den folgenden Seiten respektive Slides zu sehen. Viel Vergnügen.

Wo findet man einen

Was ist ein Was macht ein Data Scientist???



Grossunternehmen mit Mitteln für grosse Data Science Teams suchen eher

SPEZIALISTEN

Kleinere Unternehmen suchen eher

ALLROUNDER

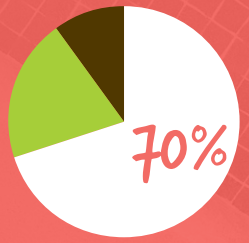
Grundsätzlich: Data Scientists sind sehr gesucht

Aus den Daten Stories machen

10% PRÄSENTIEREN

20% MODELLIERUNG

Der kreative und stark analytische Teil: Die passenden Modelle finden um die richtigen Aussagen machen zu können



70% DATENAUFBEREITUNG

Die harte Arbeit aus Mathematik, Statistik, Programmieren und Tooling zum Sichten, Zusammenführen, Vervollständigen, Vereinheitlichen und Plausibilisieren von Daten

AUSSAGE EINES DATA SCIENTISTS

«Man kann nicht einfach pauschal gute Daten haben. Man muss gute Daten für den richtigen Zweck haben.»

Data Driven Organization

«Data driven heisst auch: von den Daten lernen und glauben – und nicht einfach nach Bauchgefühl weitermachen.»

5 SCHRITTE

UM EINE DATA DRIVEN ORGANIZATION ZU WERDEN

1. Führung: Data-Analytics als zentrales und vertrauenswürdiges Element zur Entschedungsfindung anerkennen. Data informed decision making muss akzeptiert werden.
2. Ziele und Bedürfnisse der Entscheider und des Business kennen lernen um mit konkreten Zielen vorangehen zu können.
3. Hub-and-Spoke-Model: Analytics nahe/direkt im Business ansiedeln, wo die Resultate auch nützen; zusätzlich eine Globale Unit führen, welche Analytics von Bereichs-Überspannenden Themen untersucht.
4. Endbenutzer ist der Ausgangspunkt für den Aufbau von Analytics
5. Resultate messen um Optimierungspotenziale zu erkennen

Hauptrollen in einer DDO

- Data Scientist/Analyst
- Business Solution Architect
- Advanced Modelers
- Campagin Experts
- Data Visualisation Specialists
- Data Engineers

auch unterteilt in die Gruppen

- Data business people
- Data researchers
- Data developers
- Data creatives

Big Data Projekte und Prozesse

Woran Big Data Projekte scheitern

- Unklare Ziele/Ungenauer Projektumfang
- Falsche/überhöhte Erwartungen durch «Hype»
- Wenig Verständnis für Projektverzögerungen
- Viel eigene Entwicklung führt zu Labor-Modus

Entscheidend für den Erfolg von Big Data Projekten

- Klare Zielsetzung führt zu Mehrwert für Business/Kunden
- Korrektes Tooling und automatisierte Prozesse
- Abstraktions-Ebene schaffen zur Technik
- Berücksichtigung der internen Auswirungen

Für folgende Bereiche müssen wiederholbare, verwaltbare Prozesse eingeführt werden:

Zugriff
Prozess des Daten-Erhalten bis zur Speicherung der Daten

Integration
Integration und Normalisierung der Daten: Vollautomatisation anstreben

Schutz
Festlegen der Richtlinien und dynamische Maskierung von Daten zur Einhaltung derer

Daten

Analyse
Konkrete Verwendung

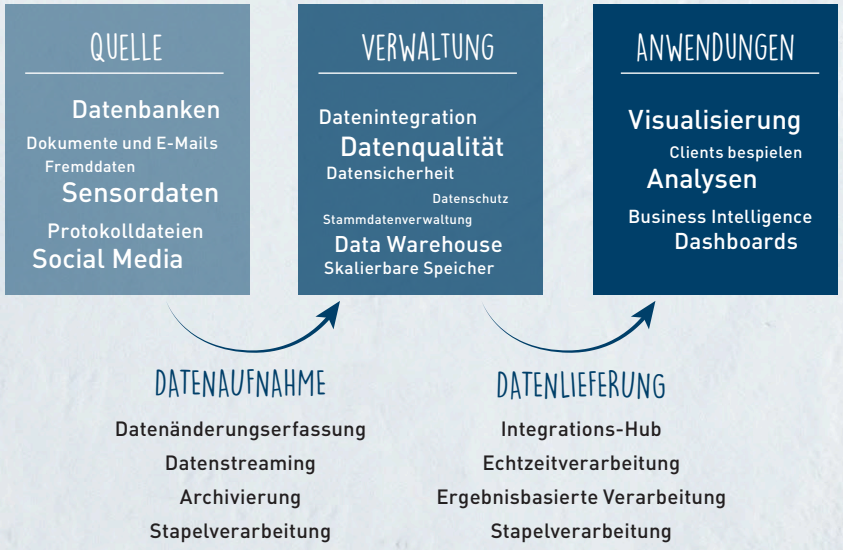
Bewältigung
Reichhaltige Sammlung konsolidierter Daten erhalten

Bereinigung
Fehler, Ungenauigkeiten und unvollständige Daten eliminieren. Prozess soll sicherstellen, dass nicht unnötig wertvolle Zeit der Analytiker hierfür verwendet wird

Operationalisierung
Erkenntnisse operationalisieren: Pipeline um zu Kommunizieren und Feedbacks einzuholen

Geschäftsanforderungen
Analyse, die oft vergessen geht. Hilft, den Fokus zu bewahren und die Auswirkungen des Big Data Projekts hoch zu halten

Die ideale Big-Data-Architektur



Pitfalls in R

für Java/PHP/JavaScript-Entwickler

Zuweisung

```
a <- 100
```

Return mit Klammern

```
return(a)
```

Arrays sind Vektoren

```
c('a','b','c')
```

Vergiss Listen, nutze Data Frames

```
data.frame(...)
```

Zusatztipp: Überschreib nicht die Funktion c

Erster Index ist nicht 0

```
first[1]
```

Schleifen vermeiden, Vektoren nutzen

```
sum(is.na(data$col))
```

Null heisst not available

```
NA
```

und das Wichtigste:

in Vektoren denken!

R - Wann man was in den Händen hält

für Java/PHP/JavaScript-Entwickler

» Debugging in R

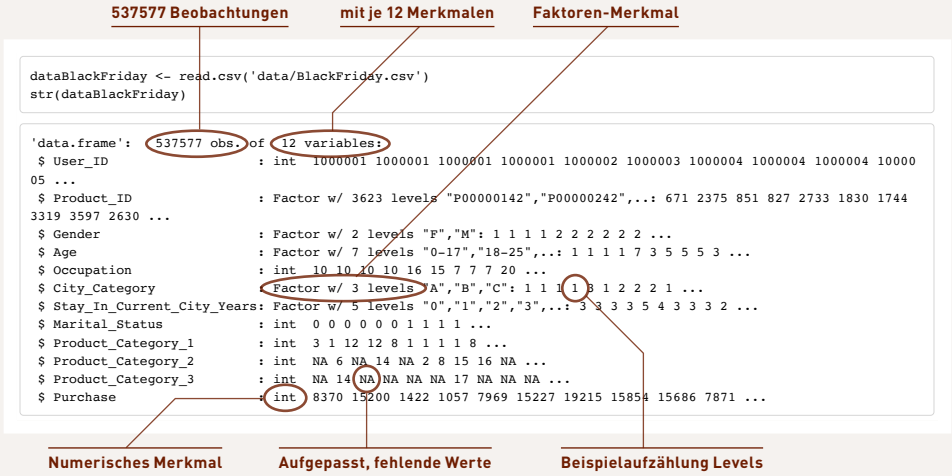
Debuggen eines Code-Abschnitts in einer RMarkdown-Datei ist in RStudio nicht möglich. Als Workaround kann man sich die zu debuggende Funktionen oder Abläufe in einer R-Script-Datei auslagern und darin debuggen.

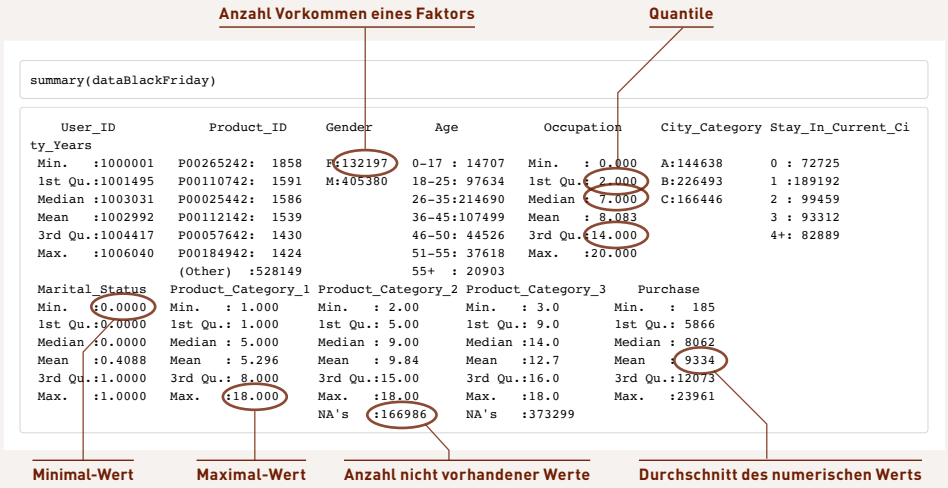
» Global Environment

In R behält man Funktionen, Werte und Datensätze ab der Zuweisung in der Globalen Umgebung. Hier gilt es beim erneuten Abspielen vom Code her sicher zu gehen, dass möglicherweise variable Werte neu zugewiesen werden. Sauberer Code ist auch beim Verteilen eines Scripts wichtig. Dazu gehören z.B. Aufrufe zum Laden von genutzen Libraries.

» str() und summary()

Ohne Debugging kann man sich mit den Befehlen str() für «Structure» und summary() eine Übersicht davon erzeugen, was für Daten vorliegen:





» Faktoren

Wenn Werte eines Vektors mit der Funktion as.character() in Zeichen umgewandelt werden können, werden sie als Faktoren dargestellt. Zum eigentlichen Wert besitzen diese ein Ebenen-Attribut (level) und ein Beschriftungs-Attribut (label).

» Konsequent mit Data Frames arbeiten

In der Erarbeitung verschiedenster Beispiele hat sich bewährt, die vorhandenen Daten wenn immer möglich als Data Frame zu halten. So kann immer gleich darauf zugegriffen werden, was gerade auch die Grafik-Herstellung mit ggplot2 stark vereinfacht.

Die Daten von JSON-Responses können mit der Funktion fromJSON() des Packages «jsonlite» einfach in R-Objekte umgewandelt werden. Bei Listen mit nur einem Wert-Typ (Spalte) wird eine Liste zurückgegeben. Bei Matrix-artigen Listen erhält man automatisch ein Data Frame.

Effektive Visualisationen

Ausdrucken war gestern.

INTERAKTIVE ANZEIGE-MÖGLICHKEITEN NUTZEN

FÜR DRILL DOWN, LABELS, GRUPPEN...

In R gibts Shiny, für grosse Visualisierungs-Tools wie tableau+ ein «Must-Have»

Daten auf den Punkt bringen.

«Ein guter Plot beantwortet stets 1 präzise Frage.»

Typische Fehler

FEHLENDE ACHSENBESCHRIFTUNGEN

manchmal auch Hinweis auf erfundene Werte

ZU VIELE DATEN UND SEKUNDÄRE DATEN

rauben die Übersicht und schwächen die Aussage

IRREFÜHRENDE SKALA

kann auch zur Manipulation eingesetzt werden

Beispiel Plot/Diagramm in ggplot2: Daten von Form getrennt

```
library(ggplot2) # Installierte Library aktivieren
econ <- read.csv(«data/economistdata.csv») # Daten lesen

d <- ggplot(econ, aes(x = CPI, y = HDI, color = Region)) + geom_point()
     |----------------- Daten-Definition -----------------|   |--- Art --|

d <- d + ggtitle("Human development vs. Corruption", subtitle="From Economist")
         |--------------- dem Plot d weiteres Element hinzufügen ---------------|

d <- d + theme(plot.title = element_text(size = rel(3.5)))
         |-- bestimmtes Theme/Theme definition anwenden --|
```

Stochastik

Statistik basics

STETIG vs. DISKRET

KONTINUIERLICH — KATEGORISCH

NUMERIC IN R — FACTOR IN R

Bei der Verwendung von Verteilungsfunktionen zuerst EINE KLARE AUFSTELLUNG VON WERTEN MACHEN:
Versuche, Erfolge, Lambda, Wahrscheinlichkeit, Erwartung

DAS GESETZ DER GROSSEN ZAHLEN

- Wenige Versuche sagen gar nichts aus
- Umso höher die Anzahl Versuche, desto stärker entspricht die Verteilung der theoretischen Wahrscheinlichkeit

KANN MAN AUCH

DAS GESETZ DER VIELEN VERSUCHE

NENNEN

Theoretische Verteilungen

NORMALVERTEILUNG
Abweichung von Nennmass, Beobachtungsfehler

BINOMIALVERTEILUNG
Urnenmodell mit Zurücklegen

HYPERGEOMETRISCHE VERTEILUNG
Urnenmodell ohne Zurücklegen

POISSONVERTEILUNG
Verteilung von seltenen Ereignissen

GLEICHVERTEILUNG
Würfeln

EXPONENTIALVERTEILUNG
Dauer von zufälligen Zeitintervallen

CHI²VERTEILUNG
Testverteilung für Hypothesentest

4 Funktionen zu jeder Verteilung

```
# am Beispiel der
# Normalverteilung

# Density
# Dichtefunktion
dnorm(x, mean, sd)

# Probability
# Wahrscheinlichkeit
pnorm(q, mean, sd)

# Quantile
qnorm(p, mean, sd)

# Random
# Erzeugung Zufallszahlen
rnorm(num, mean, sd)
```

WAHRSCHEINLICHKEIT bei diskreten Werten, Dichte bei stetigen

WAHRSCHEINLICHKEIT kummuliert

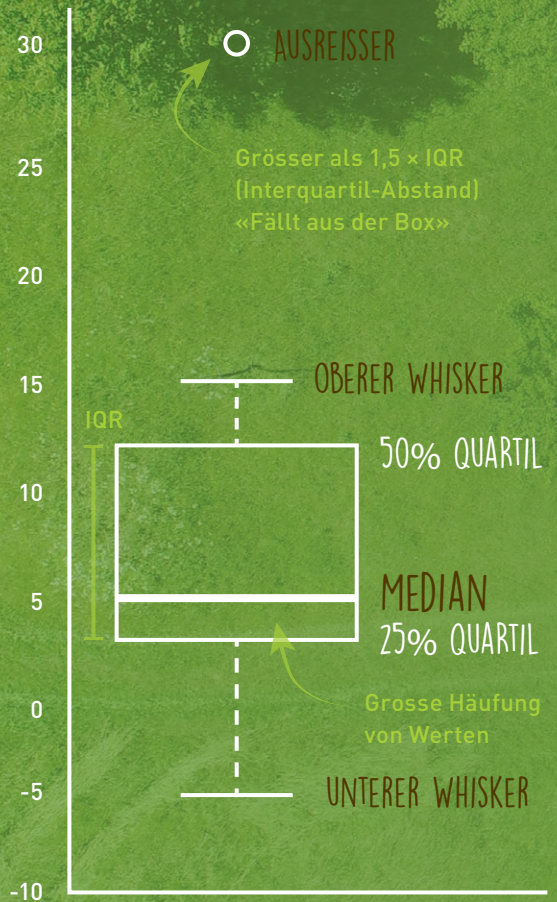
slightly more statistics

JE HÖHER DESTO BESSER
die Modellanpassung einer Linearen Regression

VARIANZ
Mittelwert der quadrierten Differenzen der gemessenen Werte zum Mittelwert

GRUNDGESAMTHEIT
Menge an Objekten, deren Merkmale untersucht werden

BESTIMMTHEITSMASS
Quadrat der Korrellation zwischen X und Y

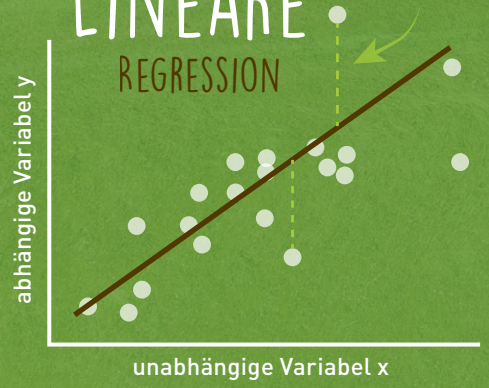


Der Boxplot erlaubt das schnelle auslesen von Ausdehnung und Verteilung

«Ein mathematisches Modell liefert immer eine Wahrscheinlichkeit.»

Vorhersagen mit Regression

LINEARE REGRESSION



STETIGE WERTE

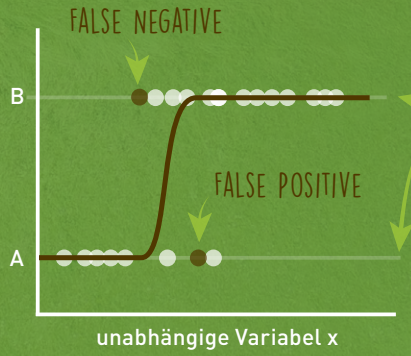
$y = a + bx$

EINFACH IN R

$lm(y \sim x_1 + x_2)$

KEINE KORRELLATION!

LOGISTISCHE REGRESSION



DISKRETE WERTE FÜR Y
NUR 2 KLASSEN

EINFACH IN R

$glm(y \sim x_1 + x_2)$

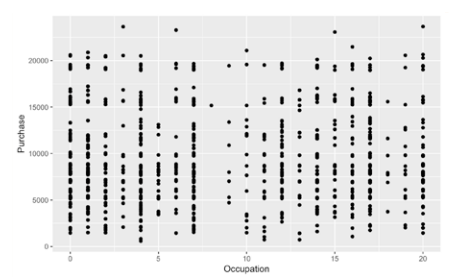
Bei Klassifizierungen würde eine linare Regression nie Sinn machen

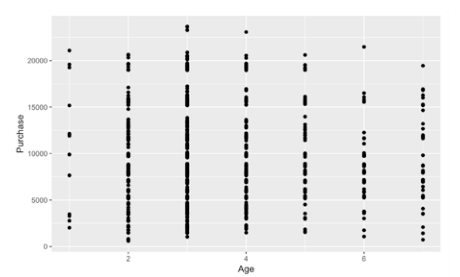
Happy Black Friday

Vorhersagen mit linearer Regession

Modell erzwingen unmöglich

Ein grosser Datensatz zum Kaufverhalten von Kunden am Black Friday sollte dazu dienen, ein Vorhersage-Modell zu erstellen. Die Merkmale wie Geschlecht, Beruf und Lebensdauer in der gleichen Stadt sind von einander unabhängig – eigentlich eine gute Ausgangslage um diese als Predictors für eine lineare Regression zu nutzen. Visualisationen wie auch ein Bestimmtheitsmass (adjusted R-Square) stieg nie über 0.12 und die Variablen-Reduktion kam auf einen hohen AIC-Wert von > 16000.





heterogene Streuung, keine ersichtliche Abhängigkeit zwischen Purchase und Merkmalen

So schön es wäre, aber aufgrund der vorhandene Merkmale war es nicht möglich, hier ein funktionierendes Modell zu erstellen.

World Happiness Ranking

Der Datensatz «World Happiness Ranking 2017» wurde halbiert. Mit der einen Hälfte wurde das Modell gebildet, mit der anderen Hälfte wurde das Resultat überprüft.
Die Variablenselektion wurde mit stepAIC durchgeführt und lieferte das folgende Resultat: Wahl von signifikanten Merkmalen und ein R^2-Wert, der gut ist und damit aussagt, wie gut die Regressions-Linie auf die Datenpunkte passt.

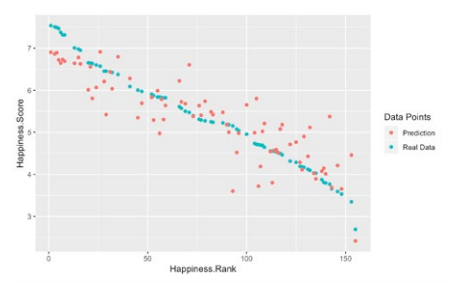
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.0662     0.2347   8.805 4.82e-13 ***
Economy..GDP.per.Capita.   1.0577     0.2684   3.940 0.000187 ***
Family                     0.8837     0.2840   3.112 0.002667 **
Health..Life.Expectancy.   1.2543     0.4671   2.685 0.008989 **
Freedom                    1.2146     0.4057   2.994 0.003770 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4656 on 72 degrees of freedom
Multiple R-squared:  0.8286,    Adjusted R-squared:  0.8191
F-statistic: 87.04 on 4 and 72 DF,  p-value: < 2.2e-16
```

Relativ hohe Signifikanz

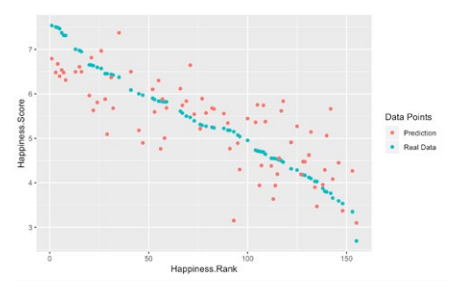
R2-Wert in akzeptablem bereich

Ein Wort zu Multikollinearität

Die für die lineare Regression gewählten Merkmale weisen untereinander alle eine relativ hohe Korrellation auf. Grundsätzlich ist das ein Zeichen auf «zu wenig Informationen», wird dadurch doch die Genauigkeit der Schätzungen nicht wirklich besser. Daher wurde ein Modell für zwei Fälle berechnet: Mit 4 und mit 1 Predictor.
Die Abweichung des Modells der 4 Predictoren weist einen Wert von 8,4% auf. Mit dem Modell mit einem einzigen, dem am stärksten korrellierenden Merkmal konnte die Prediction auf 12% genau gemacht werden. Über eine Hauptkomponentenanalyse könnte man die Merkmale zusammenführen und dadurch wohl das Modell optimieren.



Vorhersage mit Modell aus 4 Predictors:
Wirtschaft, Gesundheit, Familie und Freiheit



Vorhersage mit Modell aus 1 Predictor:
Wirtschaft

Qualität der Vorhersage und Interpretation

Mit beiden Modellen können Vorhersagen zur Glücklichkeit gemacht werden, die Aussagekraft darf aber nicht überbewertet werden. Der Ausgangsdatensatz erfasst zwar verschiedenste Bereiche des Lebens – doch gemäss dieser Analyse gibt es wohl noch weitere verborgene Parameter, die für unser aller Glück zuständig sind...

Data Warehouse vs. Data Lake

ergänzt

STRUKTURIERT

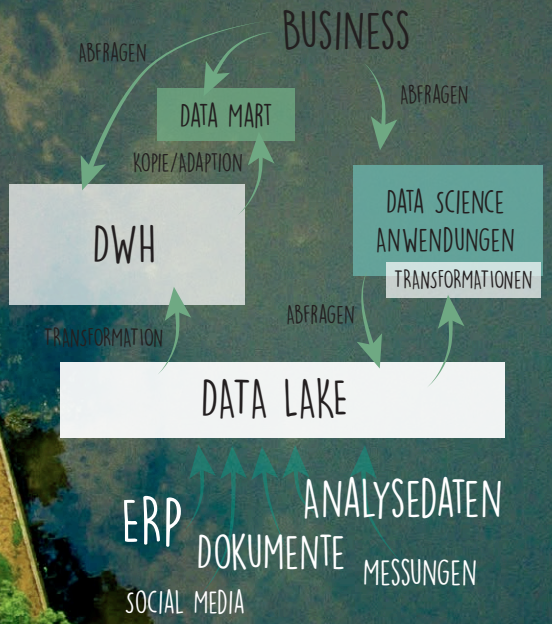
HIGH-LEVEL-ZUGANG

EINFACHE ABRUFBARKEIT

TRANSFORMATION VOR DEM IMPORT

REDUKTION VON DETAILS

METADATA-MANAGEMENT ist extrem wichtig



STRUKTUR FREI [1]

LOW-LEVEL-ZUGANG

GEZIELTE ABRUFBARKEIT

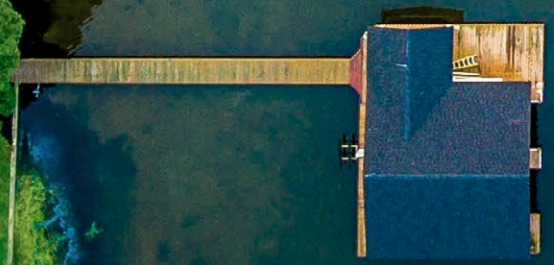
BELIEBLIGE TRANSFORMATIONEN [2] NACH DEM IMPORT

ROHDATEN: DETAILS BLEIBEN ERHALTEN

Data Mart

Ein Data Mart ist eine Kopie eines Teildaten-bestandes des Data Warehouse

Gründe für Data Marts:
- Notwendigkeit spezieller Datenstrukturen
- Performance optimieren: Verlagerung von Rechenleistungsbedarf und Zugriffen
- Zugriffsschutz/Abgrenzung von Daten



«Datensammelwut führt oft zu grosser Komplexität in der Landschaft, welche relevante Kosten mit sich bringt.» [3]

1) Ablage auf möglichst günstigem Speicherplatz
2) Eigene Tranformationen mit eigenen Modellen ermöglichen die gezielte Abrufbarkeit aus Rohdaten, enthalten aber auch das Risiko zu Wiederholung der gleichen Arbeit und unter sich inkompatiblen End-Modellen
3) Es loht sich daher, vor dem Sammeln die zu beantwortenden Hypothesen aufstellen.

Hypothesentest mit R – ein Beispiel

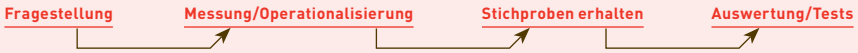
Eine provokative These überprüfen

Der Ausgangspunkt jedes Hypothesentests ist die Formulierung einer Fragestellung: der (Forschungs-)Hypothese und deren Gegenhypothese (Nullhypothese). Soll eine Hypothese anschliessend auch medial vermarktet werden, sollten solche Thesen möglichst provokativ daher kommen. Für das Beispiel wurde ein Datensatz einer Jugend-Umfrage genommen und die folgende Hypothese formuliert:

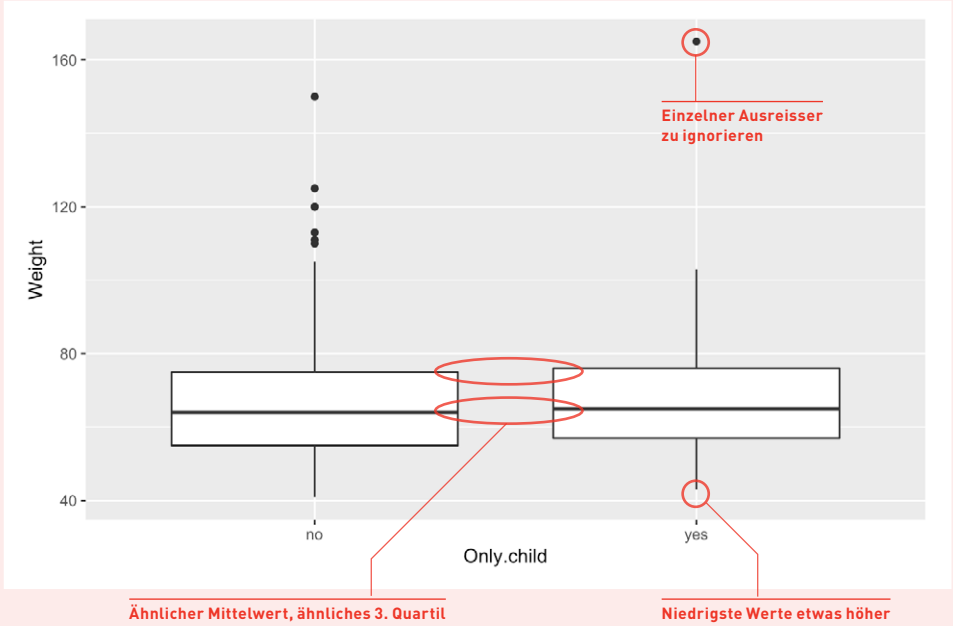
- H_1: Kinder mit Geschwistern sind dicker – die Eltern schauen schliesslich (mangels Übersicht) zu wenig auf deren Ess-Gewohnheiten!
- H_0: Kinder mit Geschwistern sind nicht schwerer als Einzelkinder.

Der Weg zum Test

Die Fragestellung markiert erst den Startpunkt. In der Regel kann man zu diesem Zeitpunkt noch auf keine Daten zugreifen. Also versucht man, das Konzept der Messung zu definieren – welche Daten wie/wo abgefragt werden könnten. Mit dem umgesetzten Konzept holt man anschliessend die Stichprobe-Daten ein. Mit diesen sind dann die Daten für den Hypothesen-Test bereit.



Hat man die Stichprobe beisammen, lohnt sich – wie so oft – eine erste Visuelle Überprüfung. Für den Fall des Vergleichs von einem kategorischen und einem kontinuierlichen Merkmal eignet sich der Boxplot ausgezeichnet.



Den richtigen Test durchführen und p-Wert finden

Die Wahl des Hypothesentests ist abhängig von den Merkmalen und von der Art der Analyse. In diesem Beispiel geht es um das Belegen eines Zusammenhangs von zwei Merkmalen, dafür eignet sich u.a. der Chi-Quadrat-Test.

```
chisq.test(x = ypsData$Only.child, y = ypsData$Weight)
Chi-Quadrat-Approximation kann inkorrekt sein

    Pearson's Chi-squared test

data:  ypsData$Only.child and ypsData$Weight
X-squared = 70.683, df = 68, p-value = 0.3882
```

Warnung aufgrund der eher kleinen Sample-Grösse

hoher p-Wert

Der daraus resultierende p-Wert muss so interpretiert werden: Wenn der p-Wert ganz klein ist, dann kann sicher immer die H_0 verworfen werden. Ist der p-Wert grösser als die gewählte Signifikanz, so kann die Null-Hypothese nicht verworfen werden. Für das hier gemachte Beispiel heisst das: Die These, dass «Kinder mit Geschwistern dicker sind», kann überhaupt nicht belegt werden.

Wozu Was sind Wie gut sind Hypothesentests??

REPRÄSENTATIVITÄT
Kriterien-Definition, die Einfluss auf den Output haben

SIGNIFIKANZ

RELEVANZ
Macht die Analyse überhaupt Sinn?

Das Management wird immer fragen:
IST DAS RESULTAT SIGNIFIKANT?
Oder hängt es rein vom Zufall ab?

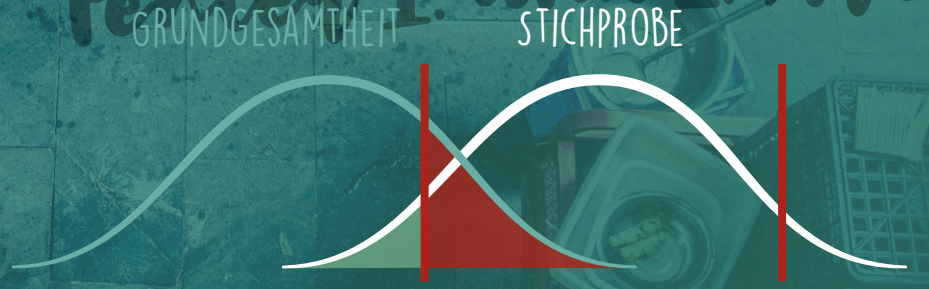
IN DER REGEL ZWISCHEN

1% MEDIZIN
2% 3% 4%
5% STANDARD
6% 7% 8% 9%
10%

Niveau

HOHE WAHRSCHEINLICHKEIT DASS H_0 VERWORFEN WIRD

Fehler 1. und 2. Art



ALPHA-FEHLER
H_0 WIRD VERWORFEN
H_0 IST RICHTIG

BETA-FEHLER
H_0 WIRD BEIBEHALTEN
H_0 IST FALSCH

Die Art der Analyse bestimmt den Test-Typ

- Zusammenhänge
  Sind zwei Variabeln von einander abhängig?
- Unterschiede
  Sind zwei oder mehr Reihen unterschiedlich?
- Überprüfung
  Entspricht die Statistik der Messung?

Dafür gibt es weit über «100 Statistical Tests»

Hypothesentests dienen dazu, um von einer Stichprobe auf die Allgemeinheit/Grundgesamtheit zu schliessen.

ERINNERN SIE SICH
noch an die Multikollinearitäts-Problematik bei der linearen Regression?
HKA ALS HILFSMITTEL

Spot on ...

... Hauptkomponenten und Faktorenanalyse

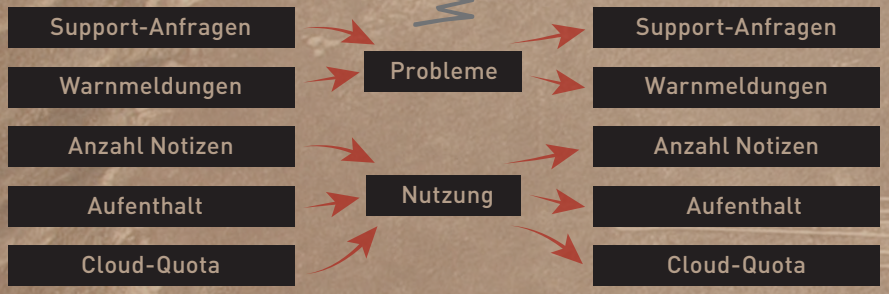
REDUKTION DER MERKMALE

Principal Components

WIRKUNG

Exploratory Factor Analysis

ÜBERPRÜFUNG DES MODELLS



Methoden in R

Vorhanden im Base-Paket
- printcomp()
- factanal()

Vorhanden im Psych-Paket
- principal()
- fa()
- fa.parallel()
- fa.diagram()
- screen()

1 **PARALLEL-ANALYSE**
Optimale Anzahl an Faktoren finden

2 **HKA/FA**
durchführen

3 **BENENNUNG**
der Faktoren

... Text Mining

Ermöglicht Auswertung von Textdaten, deren Umfang von Menschen nicht gelesen werden könnte

PANAMA PAPERS ALS BEISPIEL

IMPLIZITES WISSEN EXTRAHIEREN

ZUSAMMENHÄNGE SICHTBARMACHEN

STOPP-WÖRTER ENTFERNEN
Sie besitzen oft keine Relevanz für die Erfassung des Dokumentinhalts

Beim Aufbau von Analysen: Konzept mit einfachen Beispielen überprüfen, bei denen man die korrekte Antwort kennt

Topic Modeling im Textmining

Zurodnung von Texten zu (Themen-)Bereichen lässt sich mit Topic Modeling erreichen:
- Eine DocumentTermMatrix erstellen, aus zu analysierenden News, Büchern, Dokumenten
- Angabe wie viele Topics es entdecken soll
- Häufigste Terme pro Topic anzeigen
- Topics interpretieren – das bleibt dem Menschen überlassen!

Quasi ein Clustering

siehe nächste Seite

Zuordnungen entdecken

mit Cluster-Analysen

ENTDECKEN VON möglichst

HOMOGENEN GRUPPEN

mit möglichst GROSSER UNTERSCHEIDUNG zu anderen.

Clustering-Prozess

1 **AUSWAHL**
der geeigneten Variablen

2 **UNABHÄNGIGKEIT**
der Merkmale überprüfen und ggf. zusammenführen

3 **STANDARDISIERUNG**
der Werte

4 **ANZAHL**
Cluster definieren

5 **CLUSTER**
berechnen

6 **INTERPRETIEREN**
Kopf und Baucharbeit

2 METHODEN

um die optimale Cluster-Anzahlfür möglichst Homogene Cluster in möglichst geringer Anzahl zu finden

ELBOW-KRITERIUM

DENDOGRAMM

Anwendungsbereiche von Clusteranalysen

- Kundensegmente finden für Marketing
- Typenerkennung von Usern einer App
- Meinungsgruppen/Parteienbildung klären

Clustering liefert keine Interpretation, sondern nur Gruppen!

Songs clustern

Idee

Schon ganz zu Beginn des CAS war eine der Ideen, mit Musik/Songs etwas zu machen – gerade weil es eigentlich nicht so «greifbare» Werte gibt wie bei einer Liste mit Verkaufszahlen. Und doch: auch Musik ist «vermessbar», Anhand von Frequenzanalysen, Peak-Interpretation und anderen Methoden können Werte zu Audio-Feature-Definitionen extrahiert werden. Die Frage nun: Was lässt sich aus Audio-Features von Songs ableiten? Lassen sich Songs anhand deren eindeutig bestimmten Künstlern zuordnen?

Nutzung Spotify API

Ein Teil der Challenge bestand darin, an gute Daten zu kommen. Ein erster Ansatz, die Audio-Features selbst zu extrahieren, stellte sich als sehr aufwändig und rechenintensiv dar. Da bot sich die Nutzung der öffentlichen Spotify-REST-API an, mit den Vorteilen der umfangreichen Song-Palette und sauberer Schnittstellen-Dokumentation.

Datensatz erstellen

Nach einer manuellen Artist-ID-Suche über den Postman-Rest-Client, wurden die Top-Tracks abgerufen und beim Erhalt der Antwort bereits Datenpunkte reduziert. Mit all den Track-IDs konnten anschliessend die Audio-Features abgerufen werden.

URL zusammenstellen · REST-Aufruf · Response-Body übernehmen

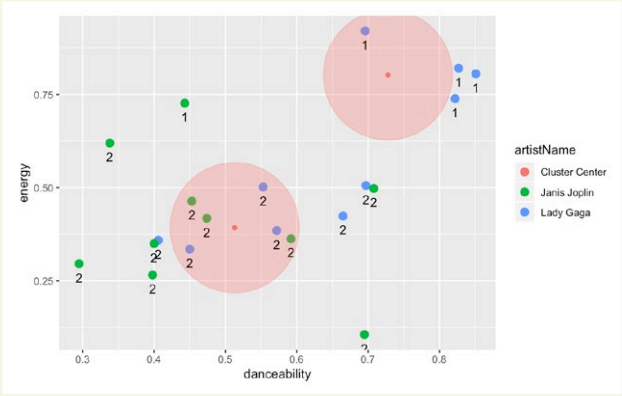
```
getTrackFeatures <- function(tracks, tokenFeatures) {
  urlFeatures <- paste("https://api.spotify.com/v1/audio-features?ids=", paste(tracks, collapse = ","), sep = "")
  log_info(urlFeatures)
  bearerFeatures <- paste("Bearer", tokenFeatures)
  respFeatures <- GET(url=urlFeatures, add_headers("Authorization"=bearerFeatures, "Content-Type"="application/x-
www-form-urlencoded"))

  featuresContent <- content(respFeatures, as = 'text')
  featuresContent <- fromJSON(featuresContent) # use parse function fromJSON
  trackFeatures <- featuresContent$audio_features # audio_features is already a dataframe
  return(trackFeatures)
}
features <- getTrackFeatures(tracks$trackId, accessJson$access_token)
```

Artists, Tracks und Features bereit: Zusammenführung anhand der IDs:

```
overall <- merge(tracks, features, by.x = 'trackId', by.y = 'id')
overall <- merge(overall, artists, by.x = 'artistId', by.y = 'artistId')
overall
```

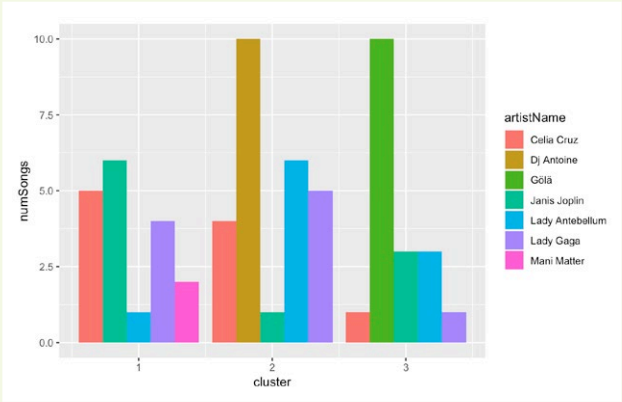
Das komplette R-Notebook unter: https://bda.chrissharkman.ch/za/Audio_Features_Clustering.nb.html

Clustering



Mit dem kompletten Datensatz konnten die Korrellationen überprüft werden, um bei stark korrellierenden auf eines der Merkmale zu verzichten. Ein erster Test dann mit nur zwei Künstlern und zwei Merkmalen (energy, danceability; Korrellation 0.32 – für diesen Kontext eine tiefe Signifikanz) auf 2 Cluster verteilt ergab dieses Bild: Tendenzen, aber keine klare Aussage. In der Idee, dass mit mehr Merkmalen ein aussagekräftigeres Resultat entstehen könnte, wurden alle Songs mit dem kmeans-Algorithmus auf optimale 3 Cluster verteilt.

Resultat



In Cluster 1 kamen eher ruhigere und melancholischere Stücke zusammen. In Cluster 2 sammelten sich die «Tanzbaren» und in Cluster 3 die «Epischen», kräftig aber nicht für die Disco.

Aussagekraft und Erfahrungen

Die Erwartungen waren sehr hoch, liessen sich aber mit den vorhandenen Daten nicht erfüllen. Eine automatische Zuordnung eines Songs zu einem Künstler war nicht «ersichtlich». Hierzu bräuchte es definitiv einen Machine Learning Ansatz: Das dann auf der nächsten Seite. Stimmungen oder «Arten von Musik» konnten aber gruppiert werden. Für einen DJ würde es sich also lohnen, mal mit «seinen» Songs Cluster zu bilden und zu schauen, wo denn die Stücke landen, bei welchen die Leute abgehen. «Big-Data-Supported-Party» wäre dann das Motto!

Was können Wozu nutzt man Machine Learning und neuronale Netzwerke

DER ANSPRUCH so funktionieren zu wollen, wie das MENSCHLICHE HIRN wird im ML/NN Bereich NICHT MEHR VERFOLGT

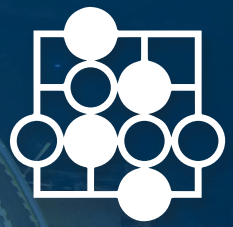
Sprachassystenten
Live-Übersetzungen

Bilderkennung
und -Klassifizierung

Betrugserkennung
in komplexen Mustern

Medizindiagnostik

Autonomes Fahren



Deep Reinforcment
Learning

Deep Dreaming

EINFACHER NEURON

Kleinste Einheit des
Neuronalen Netzwerks

OUTPUT IST
WAHRSCHEINLICHKEIT
ODER RESULTAT

0.3

«Klassifikation 2»

berechneter Wert y

$f(x) = y$

Deep-Learning Systeme könnten auch «universal approximatoren» genannt werden: sie erlernen die Approximation einer unbekannten Funktion, welche aus Input x einen Output y macht.

Warum nicht einfach Entscheidungsbäume einsetzen?

Die gute Erklärbarkeit eines Decision Trees ist der grösste Vorteil – im Vergleich zu neuronalen Netzen.

Die Nachteile sind aber zahlreich:
- geringere Qualität von Klassifikationen
- Instabilität bei wenigen «abweichenden» Daten
- Ebenso wenig Übersicht bei komplexen Situationen

Training von neuronalen Netzwerken

Das Training ist rechenintensiv. Die Ausführung der «gelernten» Formeln kann anschliessend auf schwachen Prozessoren ausgeführt werden.

EPOCHE ≠ RUN

Eine Epoche: Jeder Lern-Wert ist (mindestens) einmal «vorbeigekommen»

ANZAHL EPOCHEN

wie viele Daten brauchts?

hängt ab von
KOMPLEXITÄT DES PROBLEMS
und der
KOMPLEXITÄT DES ZU LERNENDEN ALGORITHMUS

70% als Trainingsdaten verwenden

15% als Validationsdaten gegen Overfitting im Training verwenden

15% als Testdaten verwenden

FORMEL FÜR DIE BESTE KONFIGURATION
eines neuronalen Netzwerks existiert nicht. Derzeit muss über ERFAHRUNG UND TESTS der Idealwert gefunden werden

Hyper-Parameter

ACTIVATION FUNCTIONS

SIGMOID, TANH, RELU

Entscheiden darüber, wann ein Neuron wie «feuert»

OPTIMIZER

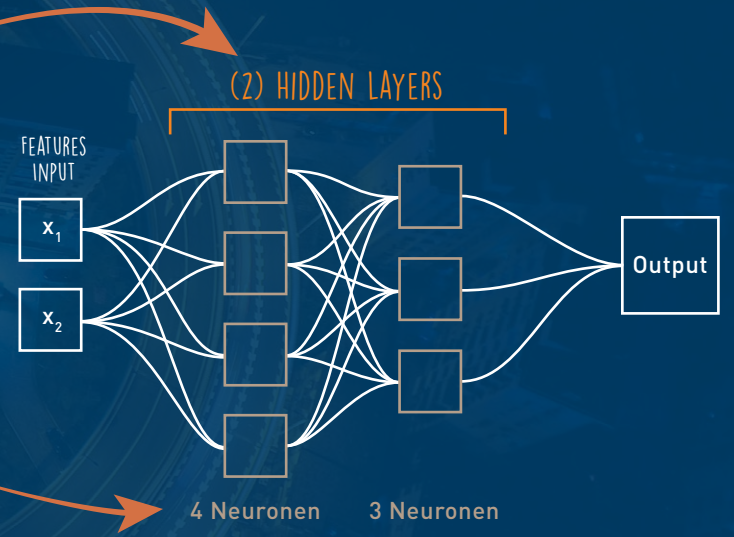
BATCH-SIZE

Nach wievielen Datensätzen die Gewichtungen angepasst werden

ANZAHL LAYER

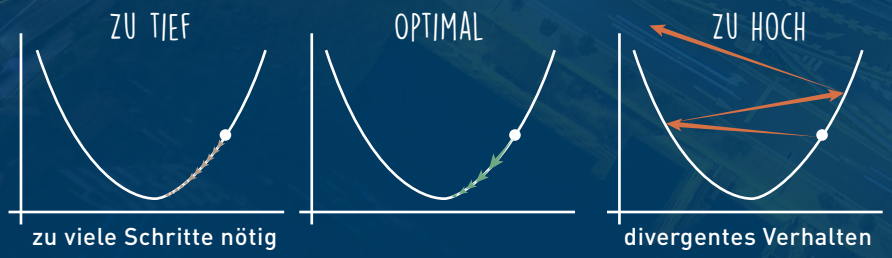
INITIALIZATION

ANZAHL UNITS PRO LAYER

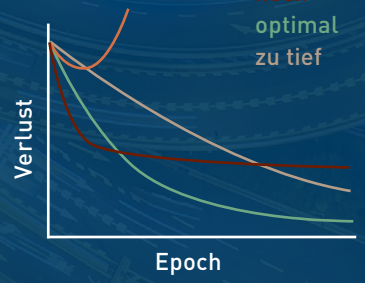


LEARNING RATE

Ob das Optimum gefunden wurde kann an der Lernkurve abgelesen werden



LERNKURVEN



Data Augmentation

Stehen einem Testset zuwenig Daten zur Verfügung und man kennt dessen Defizite, kann Data Augmentation ein Ansatz sein:
- Wird nur auf Training-Set angewendet, nicht auf Test-Set
- Wird vorallem im Bild-Bereich angewendet (Transformieren, Schneiden, Rotieren)

Fazit

Tief statt breit

Das CAS Big Data Analysis hat mich weitergebracht. Es schien zu Beginn inhaltlich überschaubar. So war es auch nicht die Breite der Themen sondern die Tiefe deren, welche die Herausforderung darstellte. Mit den ausführlich behandelten statistischen Inhalten konnte ich Bereiche abdecken, die auf meinem schulischen Weg bisher verpasst wurden. Dabei sind diese Grundlagen heute extrem wichtig, denn viele Abläufe und Fakten, welche unsere Gesellschaft steuern, werden von Systemen beeinflusst, die auf diesen Grundlagen funktionieren.

Methodische Macht

Wer die Methoden und Technologien beherrscht, dem kommt heute eine grosse Interpretations- respektive eine methodische Macht zu. Auch wenn als Grundlage wissenschaftliche Vorgaben genutzt werden, bleibt die Wahl der Merkmale oder Methoden immer auch eine persönliche Entscheidung. Auch müssen mangels Daten oft Annahmen getroffen werden, die dann als Basis dienen. Spielt man gekonnt an diesen Hebeln, kann ein Ergebnis wohl in vielen Fällen in die eine oder andere Richtung gedreht werden.
Zahlen von Auswertungen und Statistiken, ob in Berichten oder Medien, schaue ich noch einmal mit anderen Augen an. Zudem konnte ich bei Gesprächen im weiteren Umfeld feststellen, dass man in diesem Bereich durchaus «bluffen» könnte; nur ein kleiner Teil der Bevölkerung kann Angaben und Aussagen wirklich verifizieren.

Signifikant! Aha.

Sehr wichtig war auch, den Term «Signifikanz» zu verinnerlichen. Auch hier ist die Festsetzung des Signifikanz-Niveaus eine zum Teil persönliche Entscheidung des Ausführenden – und kann dadurch die definitive Aussage beeinflussen. Das Signifikanzniveau wird auch Irrtumswahrscheinlichkeit genannt. Das erklärt dessen Funktion noch besser: Damit wird bezeichnet, wie Wahrscheinlich es ist, dass eine Null-Hypothese verworfen wird, obwohl diese eigentlich richtig wäre.

Tooling

R als Tool für schnelles und dokumentiertes Skripten. Das direkte Arbeiten mit Markdown-Scripts hat sich über den ganzen Kurs hindurch bewährt, lässt es doch eine saubere Ausgabe in HTML zu, welche als «Dokumentation» sehr wertvoll ist.
Auch das «Denken in Vektoren», welches die Power von R erst ausspielen lässt, konnte sich ein bisschen etablieren. Dies dank einer grossen Anzahl an Übungsaufgaben, die im Rahmen des CAS durchgearbeitet werden konnten.

Reale Zeitinvestitionen

Die durchgespielten Beispiele wie Hypothesentest und Clustering haben gezeigt, dass die Konzeption und die Gewinnung der zur Umsetzung benötigten Daten einen Grossteil der Arbeit ausmachen. Das deckt sich mit der Arbeitsverteilung eines Data Scientisten. Die Aufwände dazu also nie unterschätzen, selbst wenn mit zusätzlicher Erfahrung effizienter gearbeitet werden kann.
Die Angewohnheit bekommen zu haben, Daten vor dem Bearbeiten einfach und schnell zu visualisieren, war ebenfalls ein grosser Gewinn. Gute Visualisierungen können ganz schnell relevantes von irrellevantem trennen. Die Mächtigkeit von simplen Visualisierungen wird aber bestimmt an vielen Orten in der Wirtschaft noch nicht ausgeschöpft. Hier gäbe es also viel zu tun – eine grosse Palette an Tools steht bereits zur Verfügung.
Die visuelle Kreationen dieser Arbeit waren letztendlich aufwändiger als zu Beginn geplant. Vorrang hatte immer der Gedanke: Beim schnellen Darüberschauen sollen die Kerninhalte des CAS sofort wieder präsent werden. So war das Ziel anstelle einer Detail-Flut eine optisch anziehende Strömung von Inhalten zu schaffen, welche an die gelesenen, gehörten und erarbeiteten Details erinnern. Damit bleibt das CAS ganz sicher in guter Erinnerung.

Einfach mehr!

Mit dem vorangehend besuchten CAS Disruptive Technologies, in dem ich die Konzepte von Machine Learning und Artificial Intelligence bereits erfahren hatte, komplettierte sich das Bild nun mit dem im Big Data Analysis vermittelten statistischen Hintergrundwissen. Mehr Wissen führt aber auch immer zu mehr fragen – so waren auch die Selflearning-Tage immer zu kurz, um in all die mögliche Literatur einzutauchen, die noch spannend gewesen wäre. Es bleibt also immer noch viel mehr zum Entdecken.

to be continued

2019 · Christian Heimann